Mahalanobis' distance disclosed in U.S Patent Number 4,991,216.

$$L_k = \sum_{i,time} B_{k_i} - 2\vec{A}^t_{k_i} \cdot \vec{X} \qquad (1)$$

where;

$$\vec{A}_k = \vec{W}^{-1} \cdot \vec{\mu}_k - \vec{W}^{-1} \cdot \vec{\mu}_x$$

$$B_k = \vec{\mu}^t_k \cdot \vec{W}^{-1} \cdot \vec{\mu}_k - \vec{\mu}^t_x \cdot \vec{W}^{-1} \cdot \vec{\mu}_x$$

$\vec{W}^{-1}$ is inverse matrix of $\vec{W}$,

$\vec{\mu}^t_k$ is transpose of a matrix of $\vec{\mu}_k$,

L_k is a distance between utterance of state (k) (phoneme order or time sequence) by a speaker and the trained pattern every category,

$\vec{\mu}_k$ is an average value of LPC cepstral coefficient vectors of state (k) (phoneme order or time sequence) every category,

$\vec{\mu}_x$ is an average value of LPC cepstral coefficient vectors of all utterances by all training speakers,

$\vec{W}$ is a covariance value of LPC cepstral coefficient vectors of all utterances by all training speakers, and

$\vec{X}$ is a continuous LPC cepstral coefficient vector of an input speech sound generated by a speaker.

Using trained patterns for categories 1, 2, and 3, distances L_{1k}, L_{2k} and L_{3k} are obtained in the following equations.

$$L_{1k} = \sum_{i,time} B_{1k_i} - 2\vec{A}^t_{1k_i} \cdot \vec{X}$$

where;

$$\vec{A}_{1k} = \vec{W}^{-1}\cdot\vec{\mu}_{1k} - \vec{W}^{-1}\cdot\vec{\mu}_x$$

$$B_{1k} = \vec{\mu}_{1k}^t\cdot\vec{W}^{-1}\cdot\vec{\mu}_{1k} - \vec{\mu}_x^t\cdot\vec{W}^{-1}\cdot\vec{\mu}_x$$

$$L_{2k} = \sum_{i,time} B_{2k_i} - 2\vec{A}_{2k_i}^t\cdot\vec{X}$$

where;

$$\vec{A}_{2k} = \vec{W}^{-1}\cdot\vec{\mu}_{2k} - \vec{W}^{-1}\cdot\vec{\mu}_x$$

$$B_{2k} = \vec{\mu}_{2k}^t\cdot\vec{W}^{-1}\cdot\vec{\mu}_{2k} - \vec{\mu}_x^t\cdot\vec{W}^{-1}\cdot\vec{\mu}_x$$

$$L_{3k} = \sum_{i,time} B_{3k_i} - 2\vec{A}_{3k_i}^t\cdot\vec{X}$$

where;

$$\vec{A}_{3k} = \vec{W}^{-1}\cdot\vec{\mu}_{3k} - \vec{W}^{-1}\cdot\vec{\mu}_x$$

$$B_{3k} = \vec{\mu}_{3k}^t\cdot\vec{W}^{-1}\cdot\vec{\mu}_{3k} - \vec{\mu}_x^t\cdot\vec{W}^{-1}\cdot\vec{\mu}_x$$

$\vec{\mu}_{1k}$ is an average value of LPC cepstral coefficient vectors of state (k) (phoneme order or time sequence) of an arrangement “Television” of speech sound elements for category 1,

$\vec{\mu}_{2k}$ is an average value of LPC cepstral coefficient vectors of state (k) (phoneme order or time sequence) of an arrangement “Television” of speech sound elements for category 2,

$\vec{\mu}_{3k}$ is an average value of LPC cepstral coefficient vectors of state (k) (phoneme order or time sequence) of an arrangement “Television” of speech sound elements for category 3,

$\vec{\mu}_x$ is an average value of LPC cepstral coefficient vectors of all utterance by all training speakers,

$\vec{W}$ is a covariance value of LPC cepstral coefficient vectors of all utterances by all training speakers, and

$\vec{X}$ is an LPC cepstral coefficient vector when a user speaks "Television".

This distance calculation uses, as an entire distribution, all utterances by speakers in various characteristic categories as discussed above. Therefore, these equations are extremely effective for the selection of trained patterns to be selected.

The present embodiment uses four words, "Television", "Video", "Air conditioner", and "Light", as defined pattern selection words. When the pattern selection words are also used for the device selection, a number of pattern selection words is preferably the same number as controlled devices. When the pattern selection and the device selection are performed using different words, a smaller number of pattern selection words can produce the same advantage. For example, when "Instruction" is used as a pattern selection word and "Instruction_Television_Increase-sound" or "Instruction_Light_Turn off" is spoken, "Television" and "Increase sound", or "Light" and "Turn off" are used as device control words. Even one pattern selection word can thus improve recognition performance of the subsequent words.

Next, distances obtained in step S403 for the trained patterns for respective categories are compared with each other (step S404). In the present embodiment, distances L_{1k}, L_{2k}, and L_{3k} obtained in step S403 are compared with each other.

Based on the comparison result in step S404, a vocabulary indicating a controlled device and a category that have the shortest distance is selected (step S405).

Patterns to be selected are reconstructed in response to the nearest pattern selected in step S405 (step S406). When the trained patterns of category 1 are selected in step S405, average values 311 of utterance by training speakers in category 1 and covariance values 321 of utterances by the training speakers in category 1 are used as the trained patterns to be selected. When the trained patterns of category 2 are selected in step S405, average values 312 of utterance by training speakers in category 2 and covariance values 322 of utterance by the training speakers in category 2 are